

Simple Complexity Analysis of Simplified Direct Search

Jakub Konečný *

Peter Richtárik †

*School of Mathematics
University of Edinburgh
United Kingdom*

September 30, 2014 (Revised[‡]: November 13, 2014)

Abstract

We consider the problem of unconstrained minimization of a smooth function in the derivative-free setting using. In particular, we propose and study a simplified variant of the direct search method (of direction type), which we call *simplified direct search (SDS)*. Unlike standard direct search methods, which depend on a large number of parameters that need to be tuned, SDS depends on a single scalar parameter only.

Despite relevant research activity in direct search methods spanning several decades, complexity guarantees—bounds on the number of function evaluations needed to find an approximate solution—were not established until very recently. In this paper we give a surprisingly *brief* and *unified analysis* of SDS for nonconvex, convex and strongly convex functions. We match the existing complexity results for direct search in their dependence on the problem dimension (n) and error tolerance (ϵ), but the overall bounds are simpler, easier to interpret, and have better dependence on other problem parameters. In particular, we show that for the set of directions formed by the standard coordinate vectors and their negatives, the number of function evaluations needed to find an ϵ -solution is $O(n^2/\epsilon)$ (resp. $O(n^2 \log(1/\epsilon))$) for the problem of minimizing a convex (resp. strongly convex) smooth function. In the nonconvex smooth case, the bound is $O(n^2/\epsilon^2)$, with the goal being the reduction of the norm of the gradient below ϵ .

Keywords: simplified direct search, derivative-free optimization, complexity analysis, positive spanning set, sufficient decrease.

*School of Mathematics, University of Edinburgh, United Kingdom (e-mail: j.konecny@sms.ed.ac.uk)

†School of Mathematics, University of Edinburgh, United Kingdom (e-mail: peter.richtarik@ed.ac.uk)
The work of both authors was supported by the Centre for Numerical Algorithms and Intelligent Software (funded by EPSRC grant EP/G036136/1 and the Scottish Funding Council). The work of J.K. was also supported by a Google European Doctoral Fellowship and the work of P.R. was supported by the EPSRC grant EP/K02325X/1.

‡In this revision (minor revision in SIAM J Opt) the main results stay the same, but we have further simplified some proofs, added some new results and further streamlined the exposition.

1 Introduction

In this work we study the problem of unconstrained minimization of a smooth function $f : \mathbb{R}^n \rightarrow \mathbb{R}$:

$$\min_{x \in \mathbb{R}^n} f(x). \quad (1)$$

We assume that we only have access to a function evaluation oracle; that is, we work in the derivative-free setting. In particular, we study a simplified variant of the direct search method of directional type, which we call *Simplified Direct Search* (SDS), and establish complexity bounds for nonconvex, convex and strongly convex objective functions f . That is, we prove bounds on the number of function evaluations which lead to the identification of an approximate solution of the optimization problem.

Despite the effort by a community of researchers spanning more than a half century [14, 23, 20, 9, 2, 15, 1, 3, 4, 22], complexity bounds for direct search have not been established until very recently in a sequence of papers by Vicente and coauthors [21, 8, 11, 13]. To the best of our knowledge, the first complexity result was established by Vicente in the case when f is smooth (and possibly nonconvex) [21], with the goal being the identification of a point x for which $\|\nabla f(x)\| \leq \epsilon$. In this work, it was shown that direct search will find such a point in $O(n^2/\epsilon^2)$ function evaluations. Subsequently, Dodangeh and Vicente [8] studied the case when f is convex (resp. strongly convex) and proved the complexity bound $O(n^2/\epsilon)$ (resp. $O(n^2 \log(1/\epsilon))$). Garmanjani and Vicente [11] established an $O(n^2/\epsilon^4)$ bound in the case when f is nonsmooth and nonconvex. Finally, Gratton, Royer, Vicente and Zhang [13] studied direct search with probabilistic descent.

1.1 Simplification of Direct Search

The direct search method, in its standard form studied in the literature (e.g., [21, 8]), depends on a large number of parameters and settings. A typical setup:

- uses 6 scalar parameters: $c > 0$ (forcing constant), $p > 1$ (exponent of the forcing function), $\alpha_0 > 0$ (initial stepsize), $0 < \beta_1 \leq \beta_2 < 1$ (stepsize decrease factors), $\gamma \geq 1$ (stepsize increase factor),
- includes a “search step” and a “poll step”,
- allows for the set of search directions (in the poll step) to change throughout the iterations (as long as their cosine measure is above a certain positive number),
- and allows the search directions to have arbitrary positive lengths.

Admittedly, this setup gives the method flexibility, as in certain situations one may wish to utilize prior knowledge or experience about the problem at hand to find a mix of parameters that works better than another mix. On the other hand, this flexibility is problematic as one has to decide on how to choose these parameters, which might be a daunting task even if one has prior experience with various parameter choices. The issue of optimal choice of parameters is not discussed in existing literature.

Let us now look at the situation from the point of view of complexity analysis. While still present in the standard form of direct search, it has been recognized that the “search step” does not influence the complexity analysis [21, 8]. Indeed, this step merely provides a plug for the

inclusion of a clever heuristic, if one is available. Moreover, direct search methods use a general forcing function¹, usually of the form $\rho(\alpha) = c\alpha^p$, where $c > 0$ is a forcing constant and $p > 1$. It can be easily inferred from the complexity results of Vicente covering the nonconvex case [21], and it is made explicit in [8], that the choice $p = 2$ gives a better bound than other choices. Still, it is customary to consider the more general setup with arbitrary p . From the complexity point of view, however, one does not need to consider the search step, nor is there any need for flexibility in the choice of p . Likewise, complexity does not improve by the inclusion of the possibility of replacing the set of search directions at every iteration as it merely depends on the smallest of the cosine measures of all these sets. In this sense, one can simply fix a single set of directions before the method is run and use that throughout.

The question that was the starting point of this paper was:

Which of the parameters and settings of direct search (in its standard form studied in the literature) are *required* from the complexity point of view? If we remove the unnecessary parameters, will it be possible to gain more insight into the workings of the method and possibly provide compact proofs leading to simpler and better bounds?

In contrast with the standard approach, SDS depends on a single parameter only (in relation to standard direct search, we fix $\beta_1 = \beta_2 = 0.5$, $p = 2$, $\gamma = 1$). As presented in Section 3, our method seems to depend on two parameters, initial stepsize $\alpha_0 > 0$ and forcing constant² $c > 0$. However, we show in Section 4 that one can, at very low cost, identify suitable α_0 or c automatically, removing the dependence on one of these parameters. Moreover, we exclude the (irrelevant-to-complexity-considerations) “search step” altogether, keep just a single set of search directions D throughout the iterative process, and keep them all of unit length (none of the extra flexibility leads to better complexity bounds). In fact, we could have gone even one step further and fixed D to be a very particular set: $D_+ = \{\pm e_i, i = 1, 2, \dots, n\}$, where e_i is the i th unit coordinate vector (or a set obtained from this by a rotation). Indeed, in all our complexity results, the choice of D enters the complexity via the fraction $|D|/\mu^2$, where μ is the cosine measure of D (we shall define this in Section 5). We conjecture that $|D|/\mu^2 \geq n^2$ for any set D for which $\mu > 0$. It can be easily seen that this lower bound is achieved, up to a factor of 2, by D_+ : $|D_+|/\mu^2 = 2n^2$. However, we still decided to keep the discussion in this paper general in terms of the set of directions D and formulate the algorithm and results that way – we believe that allowing for arbitrary directions is necessary to retain the spirit of direct search. As most optimization algorithms, including direct search, SDS also depends on the choice of an initial point x_0 .

1.2 Outline

The paper is organized as follows. We first summarize the contributions of this paper (Section 2) and then describe our algorithm (Section 3). In Section 4 we propose three initialization strategies for the method. Subsequently, in Section 5 we formulate our assumptions on the set of directions (positive spanning set). In Section 6 we state and prove our main results, three complexity theorems covering the nonconvex, convex and strongly convex cases. Finally, we conclude in Section 7.

¹Direct search accepts a new iterate only if the function value has been improved by at least $\rho(\alpha)$, where $\alpha > 0$ is the current stepsize.

²We shall see that it is optimal to choose c to be equal to the Lipschitz constant of the gradient of f . If we have some knowledge of this, which one can not assume in the derivative-free setting, it makes it easier to set c .

2 Contributions

In this section we highlight some of the contributions of this work.

1. Simplified algorithm. We study a novel variant of direct search, which we call *simplified direct search* (SDS). While SDS retains the spirit of “standard” direct search in the sense that it works with an arbitrary set of search directions forming a positive spanning set, it *depends on a single parameter* only: either the forcing constant c , or the initial stepsize parameter α_0 – depending on which of two novel initialization strategies is employed.

2. Two new initialization strategies. We describe three initialization strategies, two of which are novel and serve the purpose of automatic tuning of certain parameters of the method (either the forcing constant or the initial stepsize). The third initialization strategy is *implicitly* used by standard direct search (in fact, it is equivalent to running direct search until the first unsuccessful iteration). However, we argue that this strategy is not particularly efficient, and that it does not remove any of the parameters of the method.

3. Simple bounds. As a byproduct of the simplicity of our analysis we obtain compact and easy to interpret complexity bounds, with small constants. In Table 1 we summarize selected complexity results (bounds on the number of function evaluations) obtained in this paper. In addition to what is contained therein, we also give bounds on $\|\nabla f(x_k)\|$ in the convex and strongly convex cases, and a bound on $\|x_k - x_*\|$ in the strongly convex case – see Theorems 12 and 13.

Assumptions on f	Goal	Complexity ($c = O(1)$)	Complexity ($c = L/2$)	Thm
no additional assumptions	$\ \nabla f(x)\ < \epsilon$	$O\left(\frac{n^2 L^2 (f(x_0) - f^*)}{\epsilon^2}\right)$	$O\left(\frac{n^2 L (f(x_0) - f^*)}{\epsilon^2}\right)$	11
convex \exists minimizer x_* $R_0 < +\infty$	$f(x) - f(x_*) \leq \epsilon$	$O\left(\frac{n^2 L^2 R_0^2}{\epsilon}\right)$	$O\left(\frac{n^2 L R_0^2}{\epsilon}\right)$	12
λ -strongly convex	$f(x) - f(x_*) \leq \epsilon$	$O\left(\frac{n^2 L^2}{\lambda} \log\left(\frac{n L^2 \alpha_0^2}{\lambda \epsilon}\right)\right)$	$O\left(\frac{n^2 L}{\lambda} \log\left(\frac{n L^2 \alpha_0^2}{\lambda \epsilon}\right)\right)$	13

Table 1: Summary of the complexity results obtained in this paper.

In all cases we assume that f is L -smooth (i.e., that the gradient of f is L -Lipschitz) and bounded below by f^* ; the assumptions listed in the first column are *additional* to this. All the results in the table are referring to the setup with $D = D_+ = \{\pm e_i, i = 1, 2, \dots, n\}$. The general result is obtained by replacing the n^2 term by $|D|/\mu^2$, where μ is the cosine measure of D (however, we conjecture that the ratio $|D|/\mu^2$ can not be smaller than n^2 and hence the choice $D = D_+$ is optimal). The quantity R_0 measures the size of a specific level set of f . Definitions of all quantities appearing in the table are given in the main text.

Notice that the choice of the forcing constant c influences the complexity. It turns out that the choice $c = L/2$ minimizes the complexity bound, which then depends on L linearly. If c is a constant, the dependence becomes quadratic. Hence, the quadratic dependence on L can be

interpreted as the price for not knowing L .

4. Brief and unified analysis, better bounds. In contrast with existing results, we provide a *brief and unified* complexity analysis of direct search covering the nonconvex, convex and strongly convex cases. In particular, the proofs of our complexity theorems covering the nonconvex, convex and strongly convex cases are 6, 10 and 7 lines long, respectively, and follow the same pattern. That is, we show that in all three cases, we have the bound

$$\|\nabla f(x_k)\| \leq \frac{(\frac{L}{2} + c)\alpha_0}{\mu 2^k},$$

where $\{x_k\}$ are the “unsuccessful” iterates. We then show that in the convex case this bound implies a bound on $f(x_k) - f(x_*)$ and in the strongly convex case also on $\|x_k - x_*\|$. The difference between the three cases is that the amount of work (function evaluations) needed to reach iterate k differs as we are able to give better bounds in the convex case and even better bounds in the strongly convex case.

- **Nonconvex case.** In the nonconvex case, a relatively brief complexity analysis for direct search was already given by Vicente [21]. However, it turns out that the same analysis *simplifies substantially* when rewritten to account for the simplified setting we consider. While it is an easy exercise to check this, this simple observation was not made in the literature before. Moreover, our proof is different from this simplified proof. In more detail, the approach in [21] is to first bound the number of successful steps following the first unsuccessful step, then to bound the unsuccessful steps, and finally to bound the number of steps till the first unsuccessful step. The result is the sum of the three bounds. The complexity theorem assumes that the method converges (and a separate proof is needed for that). In contrast, we do not require a convergence proof to proceed to the complexity proof. Also, we show that in SDS it is the number of “unsuccessful” steps which directly determine the quality of the solution; while the number of “successful steps” provides a bound on the workload (i.e., number of function evaluations) needed to find that solution.
- **Convex and strongly convex case.** Existing analysis in the convex and strongly convex cases seems to be more involved and longer [8, pages 8-18] than the analysis in the nonconvex case [21]. However, we observe the analysis and results in [8] would simplify in our simplified setting (i.e., for SDS). Still, the complexity results are weaker than our bounds. For instance, Theorem 4.1 in [8], giving bounds on function values in the convex case, has *quadratic dependence* on R_0 (we have linear dependence). We should also remark that in some sense, the setting in [8] is already simplified; in particular, the analysis in [8] only works under the *assumption* that the stepsizes of the method do not grow above a certain constant M . Note that SDS removes stepsize increases altogether.

5. Extensions. The simplicity of the method and of the analysis makes it easier to propose extensions and improvements. For instance, one may wish to study the complexity of SDS for different function classes (e.g., convex and nonsmooth), different optimization problem (e.g., stochastic optimization, constrained optimization) and for variants of SDS (e.g., by introducing randomization).

6. Wider context. We now briefly review some of the vast body of literature on derivative-free optimization.

It is well known [16, Section 1.2.3] that for the problem of unconstrained minimization of a smooth (and not necessarily convex) function, gradient descent takes at most $\mathcal{O}(1/\epsilon^2)$ iterations to drive the norm of the gradient below ϵ . Such a bound has been proved tight in [5]. In the context of derivative-free methods, Nesterov’s random Gaussian approach [17] attains the complexity bound $\mathcal{O}(n^2/\epsilon^2)$. Vicente matches this result with a (deterministic) direct search algorithm [21], and so does our analysis of direct search. Cartis et al. [6] derived a bound of $\mathcal{O}(n^2/\epsilon^{3/2})$ for a variant of their adaptive cubic overestimation algorithm using finite differences to approximate derivatives. In this setting, Ghadimi and Lan [12] achieve better (linear) dependence on n by considering a slightly more special class of problems.

In the convex case, gradient descent achieves the improved bound of $\mathcal{O}(1/\epsilon)$ [16, Section 2.1.5]. For derivative-free methods, this rate is also achievable by Nesterov’s random Gaussian approach [17] ($\mathcal{O}(n/\epsilon)$) and by direct search [8] ($\mathcal{O}(n^2/\epsilon)$). We match the latter result in this paper.

Optimal (also known as accelerated/fast) gradient methods employ a two step strategy, and enjoy the complexity bound of $\mathcal{O}(1/\epsilon^{1/2})$ iterations [16, Section 2.2.1]. The derivative-free analogue of this method, also developed by Nesterov [17], needs $\mathcal{O}(n/\epsilon^{1/2})$ function evaluations. To the best of our knowledge, there are no results on (non-randomized) direct search methods that would attain this bound.

In the strongly convex setting, gradient descent achieves linear convergence, i.e., the bound on number of iterations is $\mathcal{O}(\log(1/\epsilon))$. This rate is also achievable in derivative-free setting by multiple methods [8, 17, 7], including our version of direct search.

A recent work of Recht et al. [18] goes beyond the zero-order oracle. Central in their work is a pairwise comparison oracle, that returns only the order of function values at two different points. They provide lower and upper complexity bounds for both deterministic and stochastic oracles. A related randomized coordinate descent algorithm is proposed, that also achieves $\mathcal{O}(n \log(1/\epsilon))$ calls of the oracle for strongly convex functions. Duchi et al. [10] prove tight bounds for online bandit convex optimization problems with multi-point feedback. However, the optimal iteration complexity for single point evaluation still remains an open problem. Yet another related approach, where one has access to partial derivatives, is the randomized coordinate descent method [19]. The iteration complexity of the method is $\mathcal{O}(n/\epsilon)$ in the convex case and $\mathcal{O}(n \log(1/\epsilon))$ in the strongly convex case. This method can be extended to the derivative-free setting by considering finite difference approximation of partial derivatives.

3 Simplified Direct Search

In Section 3.1 we describe the Simplified Direct Search (SDS) method – in a user-friendly notation (Algorithm 1). Subsequently, in Section 3.2 we rewrite the method into an analysis-friendly notation (Algorithm 2) and collect in a lemma a few elementary observations which will be useful later.

3.1 The method: user-friendly notation

SDS (Algorithm 1) works with a fixed finite set D of vectors in \mathbb{R}^n forming a positive spanning set (we shall discuss this in detail in Section 5). The method is only allowed to take steps of positive lengths, along directions $d \in D$. That is, every update step is of the form

$$x \leftarrow x + \alpha d, \tag{2}$$

for some stepsize $\alpha > 0$ and $d \in D$. Updates of this form are repeated while they lead to a function decrease of at least $c\alpha^2$ (these steps are called “successful steps” and the decrease is called “sufficient decrease”), where $c > 0$ is a “forcing constant” that remains fixed throughout the iterative process. That is, we move from x to $x + \alpha d$ if $d \in D$ is found for which $f(x + \alpha d) \leq f(x) - c\alpha^2$.

Algorithm 1 Simplified Direct Search (SDS): user-friendly form

1. INPUT: starting point $x \in \mathbb{R}^n$; stepsize $\alpha > 0$; forcing constant $c > 0$; finite set $D \subset \mathbb{R}^n$
 2. Repeat
 - $\alpha \leftarrow \frac{1}{2}\alpha$
 - **while** there exists $d \in D$ such that $f(x + \alpha d) \leq f(x) - c\alpha^2$
 set $x \leftarrow x + \alpha d$
-

There are several ways how, given a stepsize α , a vector $d \in D$ leading to sufficient decrease can be found; the method does not prescribe this in more detail. For instance, one may simply order the directions in D and search through them one by one, accepting the *first one* that leads to sufficient decrease. Or, one may search through all directions, and if more lead to sufficient decrease, pick the *best one*. Needless to say, the search for a direction can be easily *parallelized*. In summary, the method (and our analysis) is agnostic about the way the selection of direction $d \in D$ leading to sufficient decrease is implemented.

Once no step of the form (2) leads to sufficient decrease, i.e., if $f(x + \alpha d) > f(x) - c\alpha^2$ for all $d \in D$ (we call such steps “unsuccessful”), we do not³ update x , halve⁴ the stepsize ($\alpha \leftarrow \alpha/2$) and repeat the process. The rationale behind halving the stepsize also at the beginning of the method will be described in Section 4.

Note that the method is monotonic, i.e., at each successful step the function value decreases, while it stays the same at unsuccessful steps.

3.2 The method: analysis-friendly notation

As described above, the method is conceptually very simple. However, for the sake of analysis, it will be useful to establish notation in which we give a special label, $\{x_k\}$, to the unsuccessful iterates x , i.e., to points x for which no point of the form $x + \alpha d$, $d \in D$, where α is the current stepsize, leads to sufficient decrease. The reason for this is the following: we will prove quality-of-solution guarantees for the points x_k (in particular, we shall show that $\|\nabla f(x_k)\| = O(1/2^k)$), whereas the number of successful points will determine the associated cost (i.e., number of function evaluations) of achieving this guarantee.

³In practice, one would update x to the best of the points $x + \alpha d$, $d \in D$, if any of them has a smaller function value than x . Under such a modification, our theoretical results would either be unchanged (in the convex case: since here we give guarantees in terms of the function value – which can only get better this way), or would only need minor modifications (in the nonconvex case: here we give guarantees on the norm of the gradient).

⁴We have chosen, for simplicity of exposition, to present the algorithm and the analysis in the case when the stepsize is divided by 2 at each unsuccessful step. However, one can replace the constant 2 by any other constant larger than 1 and all the results of this paper hold with only minor and straightforward modifications.

With this in mind, Algorithm 1 can be rewritten into an analysis-friendly form, obtaining Algorithm 2. For convenience, let us now describe the method again, using the new notation.

Algorithm 2 Simplified Direct Search (SDS): analysis-friendly form

1. INPUT: starting point $x_0 \in \mathbb{R}^n$; stepsize $\alpha_0 > 0$; forcing constant $c > 0$; finite set $D \subset \mathbb{R}^n$
2. For $k \geq 1$ repeat
 - Set $x_k^0 = x_{k-1}$ and $\alpha_k = \frac{1}{2}\alpha_{k-1}$
 - Let $x_k^0, \dots, x_k^{l_k}$ be generated by

$$x_k^{l+1} = x_k^l + \alpha_k d_k^l, \quad d_k^l \in D, \quad l = 0, \dots, l_k - 1,$$

so that the following relations hold:

$$f(x_k^{l+1}) \leq f(x_k^l) - c\alpha_k^2, \quad l = 0, \dots, l_k - 1, \quad (3)$$

and

$$f(x_k^{l_k} + \alpha_k d) > f(x_k^{l_k}) - c\alpha_k^2 \quad \text{for all } d \in D. \quad (4)$$

- Set $x_k = x_k^{l_k}$
-

We start⁵ with an initial iterate $x_0 \in \mathbb{R}^n$, an initial stepsize parameter $\alpha_0 > 0$ and a forcing constant $c > 0$. Given x_{k-1} and α_{k-1} , we seek to determine the next iterate x_k . This is done as follows. First, we initialize our search for x_k by setting $x_k^0 = x_{k-1}$ and decrease the stepsize parameter: $\alpha_k = \frac{\alpha_{k-1}}{2}$. Having done that, we try to find $d \in D$ for which the following sufficient decrease condition holds:

$$f(x_k^0 + \alpha_k d) \leq f(x_k^0) - c\alpha_k^2.$$

If such d exists, we call it d_k^0 , declare the search step *successful* and let $x_k^1 = x_k^0 + \alpha_k d_k^0$. Note that the identification of x_k^1 requires, in the worst case, $|D|$ function evaluations (assuming $f(x_0)$ was already computed before). This process is repeated until we are no longer able to find a successful step; that is, until we find $x_k^{l_k}$ which satisfies (4). Such a point must exist if we assume that f is bounded below.

Assumption 1 (Boundedness of f). *f is bounded below. That is, $f^* \stackrel{\text{def}}{=} \inf\{f(x) : x \in \mathbb{R}^n\} > -\infty$.*

Indeed, under Assumption 1 it is not possible to keep decreasing the function value by the positive constant $c\alpha_k^2$, and hence l_k must be finite, and $x_k = x_k^{l_k}$ is well defined. In particular, from (3) we can see that

$$f(x_k) = f(x_k^{l_k}) \leq f(x_k^{l_k-1}) - c\alpha_k^2 \leq f(x_k^0) - l_k c\alpha_k^2 = f(x_{k-1}) - l_k c\alpha_k^2,$$

⁵In Section 4 we show that it is possible to remove the dependence of the method on α_0 or c ; hence, SDS depends on a single parameter only: (the method still depends on x_0 and D , but these could simply be set to $x_0 = 0$ and $D = D_+ = \{\pm e_i, i = 1, 2, \dots, n\}$). We prefer the slightly more general form which gives freedom to the user in choosing x_0 and D .

from which we obtain the following bound:

$$l_k \leq \frac{f(x_{k-1}) - f(x_k)}{c\alpha_k^2} \leq \frac{f(x_0) - f^*}{c\alpha_k^2} = \frac{4^k(f(x_0) - f^*)}{c\alpha_0^2}. \quad (5)$$

This way, we produce the sequence

$$x_k^0, x_k^1, \dots, x_k^{l_k},$$

set $x_k = x_k^{l_k}$, and proceed to the next iteration.

Note also that it is possible for l_k to be equal to 0, in which case we have $x_k = x_{k-1}$. However, there is still progress, as the method has learned that the stepsize α_k does not lead to a successful step.

We now summarize the elementary observations made above.

Lemma 2. *Let Assumption 1 be satisfied. Then*

(i) *The counters l_k are all finite, bounded as in (5), and hence the method produces an infinite sequence of iterates $\{x_k\}_{k \geq 0}$ with non-increasing function values: $f(x_{k+1}) \leq f(x_k)$ for all $k \geq 0$.*

(ii) *The total number of function evaluations up to iteration k is bounded above by*

$$N(k) \stackrel{\text{def}}{=} 1 + \sum_{j=1}^k |D|(l_j + 1). \quad (6)$$

(iii) *For all $k \geq 1$ we have*

$$f(x_k + \alpha_k d) > f(x_k) - c\alpha_k^2 \quad \text{for all } d \in D. \quad (7)$$

Proof. We have established part (i) in the preceding text. The leading “1” in (6) is for the evaluation of $f(x_0)$. Having computed $f(x_{j-1})$, the method needs to perform at most $|D|(l_j + 1)$ function evaluations to identify x_j : up to $|D|$ evaluations to identify each of the points $x_j^1, \dots, x_j^{l_j}$ and further $|D|$ evaluations to verify that (4) holds. It remains to add these up; which gives (ii). Part (iii) follows trivially by inspecting (4). \square

4 Initialization

It will be convenient to assume—without loss of generality as we shall see—that (7) holds for $k = 0$ also. In fact, this is the reason for halving the initial stepsize at the *beginning* of the iterative process. Let us formalize this as an assumption:

Assumption 3 (Initialization). *The triple (x_0, α_0, c) satisfies the following relation:*

$$f(x_0 + \alpha_0 d) > f(x_0) - c\alpha_0^2 \quad \text{for all } d \in D. \quad (8)$$

In this section we describe three ways of identifying a triple (x_0, α_0, c) for which Assumption 3 is satisfied; some more efficient than others.

- *Bootstrapping initialization.* Finds a suitable starting point x_0 .
- *Stepsize initialization.* Finds a “large enough” stepsize α_0 .
- *Forcing constant initialization.* Finds a “large enough” forcing constant c .

We will now show that initialization can be performed very efficiently (in particular, we prefer *stepsize initialization* and forcing constant initialization to bootstrapping), and hence Assumption 3 does not pose any practical issues. In fact, we found that stepsize initialization has the capacity to dramatically improve the practical performance of SDS, especially when compared to the bootstrapping strategy — which is the strategy *implicitly* employed by direct search methods in the literature [21, 8]. Indeed, bootstrapping initialization is equivalent to running SDS without any initialization whatsoever.

4.1 Bootstrapping initialization

A natural way to initialize—although this may be very inefficient in both theory and practice—is to simply run the direct search method itself (that is, Algorithm 1; without halving the stepsize), using any triple $(\tilde{x}_0, \alpha_0, c)$, and stop as soon as the first unsuccessful iterate x is found. We can then set $x_0 \leftarrow x$ and keep α_0 and c . This method is formally described in Algorithm 3.

Algorithm 3 Bootstrapping initialization (finding suitable x_0)

1. INPUT: $\tilde{x}_0 \in \mathbb{R}^n$; stepsize $\alpha_0 > 0$; forcing constant $c > 0$; $D \subset \mathbb{R}^n$
 2. $x \leftarrow \tilde{x}_0$
 3. **while** there exists $d \in D$ such that $f(x + \alpha_0 d) \leq f(x) - c\alpha_0^2$
 - set $x \leftarrow x + \alpha_0 d$
 4. OUTPUT: $x_0 = x$, $\alpha_0 = \alpha_0$ and $c = c$
-

Lemma 4. *Let Assumption 1 (f is bounded below) be satisfied. Then Algorithm 3 outputs triple (x_0, α_0, c) satisfying Assumption 3. Its complexity is*

$$I_{x_0} \stackrel{\text{def}}{=} |D| \left(\frac{f(\tilde{x}_0) - f^*}{c\alpha_0^2} + 1 \right) \quad (9)$$

function evaluations (not counting the evaluation of $f(x_0)$).

Proof. Starting from $f(\tilde{x}_0)$, the function value cannot decrease by more than $f(\tilde{x}_0) - f^*$, hence, x is updated at most $(f(\tilde{x}_0) - f^*)/(c\alpha_0^2)$ times. During each such decrease at most $|D|$ function evaluations are made. There is an additional $|D|$ term to account for checking at the end that (8) holds. [Note that we are using the same logic which led to bound (5).] \square

4.2 Stepsize initialization

We now propose a much more efficient initialization procedure (Algorithm 4). Starting with $(x_0, \tilde{\alpha}_0, c)$, the procedure finds large enough α_0 (in particular, $\alpha_0 \geq \tilde{\alpha}_0$), so that the triple (x_0, α_0, c) satisfies Assumption 3 – if f is *convex*. It turns out, as we shall see, that the initialization step is not needed in the nonconvex case, and hence this is sufficient. At the same time, we would like to be able to use this initialization algorithm also in the nonconvex case – simply because in derivative-free optimization we may not *know* whether the function we are minimizing is or is not convex! So, in this case we would like to be at least able to say that the initialization algorithm stops after a certain (small) number of function evaluations, and be able to say something about how large α_0 is.

Algorithm 4 Stepsize initialization (finding large enough α_0)

1. INPUT: $x_0 \in \mathbb{R}^n$; stepsize $\tilde{\alpha}_0 > 0$; forcing constant $c > 0$; $D = \{d_1, d_2, \dots, d_p\}$
 2. $i \leftarrow 1$ and $\alpha \leftarrow \tilde{\alpha}$
 3. **while** $i \leq |D|$
 - **if** $f(x_0 + \alpha d_i) \leq f(x_0) - c\alpha^2$ **then** set $\alpha \leftarrow 2\alpha$
 - **else** $i \leftarrow i + 1$
 4. OUTPUT: $\alpha_0 = \alpha$
-

The following result describes the behaviour of the stepsize initialization method.

Lemma 5. *Let Assumption 1 (f is bounded below) be satisfied.*

(i) *Algorithm 4 outputs α_0 satisfying*

$$1 \leq \frac{\alpha_0}{\tilde{\alpha}_0} \leq \max \left\{ 1, 2\sqrt{\frac{f(x_0) - f^*}{c\tilde{\alpha}_0^2}} \right\} \stackrel{\text{def}}{=} M,$$

and performs in total at most

$$I_{\alpha_0} \stackrel{\text{def}}{=} |D| + \log_2 M = |D| + \max \left\{ 0, 1 + \frac{1}{2} \log_2 \left(\frac{f(x_0) - f^*}{c\tilde{\alpha}_0^2} \right) \right\} \quad (10)$$

function evaluations (not counting the evaluation of $f(x_0)$).

(ii) *If f is convex, then the triple (x_0, α_0, c) satisfies Assumption 3.*

Proof. (i) If at some point during the execution of the algorithm we have $\alpha > \sqrt{(f(x_0) - f^*)/c} \stackrel{\text{def}}{=} h$, then the “if” condition cannot be satisfied, and hence α will not be further doubled. So, if $\tilde{\alpha} \leq h$, then $\alpha \leq 2h$ for all α generated throughout the algorithm, and if $\tilde{\alpha} > h$, then $\alpha = \tilde{\alpha}$ throughout the algorithm. Consequently, $\alpha \leq \max\{\tilde{\alpha}, 2h\}$ throughout the algorithm.

Now note that at each step of the method, either α is doubled, or i is increased by one. Since, as we have just shown, α remains bounded, and D is finite, the algorithm will stop. Moreover,

the method performs function evaluation of the form $f(x_0 + \alpha d_i)$, where α can assume at most $1 + \log_2 M$ different values and d_i at most $|D|$ different values, in a fixed order. Hence, the method performs at most $|D| + \log_2 M$ function evaluations (not counting $f(x_0)$).

(ii) Note that for each $d_i \in D$ there exists $\alpha_i \leq \alpha_0$ for which

$$f(x_0 + \alpha_i d_i) > f(x_0) - c\alpha_i^2. \quad (11)$$

Indeed, this holds for α_i equal to the value of α at the moment when the index i is increased. We now claim that, necessarily, inequality (11) must hold with α_i replaced by α_0 . We shall show that this follows from convexity. Indeed, by convexity,

$$\frac{f(x_0 + \alpha_0 d_i) - f(x_0)}{\alpha_0} \geq \frac{f(x_0 + \alpha_i d_i) - f(x_0)}{\alpha_i},$$

which implies that

$$f(x_0 + \alpha_0 d_i) \geq f(x_0) + (f(x_0 + \alpha_i d_i) - f(x_0)) \frac{\alpha_0}{\alpha_i} > f(x_0) - c\alpha_i^2 \frac{\alpha_0}{\alpha_i} \geq f(x_0) - c\alpha_0^2.$$

We have now established (8), and hence the second statement of the theorem is proved. \square

The runtime of the stepsize initialization method (Algorithm 4), given by (10), is negligible compared to the runtime of the bootstrapping initialization method (Algorithm 3).

4.3 Forcing constant initialization

In Algorithm 5 we set c to a large enough value, given x_0 and α_0 .

Algorithm 5 Forcing constant initialization (finding suitable c)

1. INPUT: $x_0 \in \mathbb{R}^n$; stepsize $\alpha_0 > 0$; $D \subset \mathbb{R}^n$
 2. OUTPUT: $x_0 = x_0$, $\alpha_0 = \alpha_0$ and $c = 1 + \max \left\{ 0, \frac{f(x_0) - \min_{d \in D} f(x_0 + \alpha_0 d)}{\alpha_0^2} \right\}$
-

Lemma 6. *Algorithm 5 outputs triple (x_0, α_0, c) satisfying Assumption 3. Its complexity is*

$$I_c \stackrel{\text{def}}{=} |D| \quad (12)$$

function evaluations (not counting the evaluation of $f(x_0)$).

Proof. By construction, c is positive and $c\alpha_0^2 > f(x_0) - f(x_0 + \alpha_0 d)$ for all $d \in D$. The method needs to evaluate $f(x_0 + \alpha_0 d)$ for all $d \in D$. \square

5 Positive spanning sets and their cosine measure

Clearly, the set of directions, D , needs to be rich enough so that every point in \mathbb{R}^n (in particular, the optimal point) is potentially reachable by a sequence of steps of SDS. In particular, we will assume that D is a *positive spanning set*; that is, the conic hull of D is \mathbb{R}^n :

$$\mathbb{R}^n = \left\{ \sum_i t_i d_i, d_i \in D, t_i \geq 0 \right\}.$$

Proposition 7 lists several equivalent characterizations of a positive spanning set. We do not need the result to prove our complexity bounds; we include it for the benefit of the reader. A proof sketch can be found in the appendix.

Proposition 7. *Let D be a finite set of nonzero vectors. The following statements are equivalent:*

- (i) D is a positive spanning set.
- (ii) The cosine measure of D , defined below⁶, is positive:

$$\mu(D) = \mu \stackrel{\text{def}}{=} \min_{0 \neq v \in \mathbb{R}^n} \max_{d \in D} \frac{\langle v, d \rangle}{\|v\| \|d\|} > 0. \quad (13)$$

Above, $\langle \cdot, \cdot \rangle$ is the standard Euclidean inner product and $\|\cdot\|$ is the standard Euclidean norm.

- (iii) The convex hull of D contains 0 in its interior⁷.

This is formalized in the following assumption.

Assumption 8 (Positive spanning set). *D is a finite set of unit-norm in \mathbb{R}^n forming a positive spanning set.*

Note that we assume that all vectors in D are of unit length. While the algorithm and theory can be extended in a straightforward way to allow for vectors of different lengths (which, in fact, is standard in the literature), this does not lead to an improvement in the complexity bounds and merely makes the analysis and results a bit less transparent. Hence, the unit length assumption is enforced for convenience.

This assumption is standard in the literature on direct search. Indeed, it is clearly necessary as otherwise it is not possible to guarantee that any point (and, in particular, the optimal point) can be reached by a sequence of steps of the algorithm.

The cosine measure μ has a straightforward geometric interpretation: for each nonzero vector v , let $d \in D$ be the vector forming the smallest angle with v and let $\mu(v)$ be the cosine of this angle. Then $\mu = \min_v \mu(v)$. That is, for every nonzero vector v there exists $d \in D$ such that the cosine of the angle between these two vectors is at least $\mu > 0$ (i.e., the angle is acute). In the analysis, we shall consider the vector v to be the negative gradient of f at the current point. While this gradient is unknown, we know that there is a direction in D which approximates it well, with the size of μ being a measure of the quality of that approximation: the larger μ is, the better.

⁶Note that the continuous function $v \mapsto \max_{d \in D} \langle v, d \rangle / \|d\|$ attains its minimizer on the compact set $\{v : \|v\| = 1\}$. Hence, it is justified to write minimum in (13) instead of infimum.

⁷It is clear from the proof provided in the appendix that if (ii) holds, then the convex hull of $\{d/\|d\| \mid d \in D\}$ contains the ball centered at the origin of radius μ .

Equivalently, μ can be seen as the largest scalar such that for all nonzero v there exists $d \in D$ so that the following inequality holds:

$$\mu \|v\| \|d\| \leq \langle v, d \rangle. \quad (14)$$

This is a reverse of the Cauchy-Schwarz inequality, and hence, necessarily, $\mu \leq 1$. However, for $\mu = 1$ to hold we would need D to be dense on the unit sphere. For better insight, consider the following example. If D is chosen to be the “maximal positive basis” (composed of the coordinate vectors together with their negatives: $D = \{\pm e_i \mid i = 1, \dots, n\}$), then

$$\mu = \frac{1}{\sqrt{n}}. \quad (15)$$

6 Complexity analysis

In this section we state and prove our main results: three complexity theorems covering the non-convex, convex and strongly convex case. We also provide a brief discussion.

In all results of this section we will make the following assumption.

Assumption 9 (*L-smoothness of f*). f is L -smooth. That is, f has a Lipschitz continuous gradient, with a positive Lipschitz constant L :

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \text{for all } x, y \in \mathbb{R}^n. \quad (16)$$

6.1 Key lemma

In this section we establish a key result that will drive the analysis. The result is standard in the analysis of direct search methods [7], although we only need it in a simplified form⁸. Moreover, we will use it in a novel way, which leads to a significant simplification (especially in the convex cases) and unification of the analysis, and to sharper and cleaner complexity bounds. We include the proof for completeness.

Lemma 10. *Let Assumption 9 (f is L -smooth) and Assumption 8 (D is a positive spanning set) be satisfied and let $x \in \mathbb{R}^n$ and $\alpha > 0$. If $f(x) - c\alpha^2 < f(x + \alpha d)$ for all $d \in D$, then*

$$\|\nabla f(x)\| \leq \frac{1}{\mu} \left(\frac{L}{2} + c \right) \alpha. \quad (17)$$

⁸Lemma 10 is usually stated in a setting with the vectors in D allowed to be of arbitrary lengths, and with $c\alpha^2$ replaced by an arbitrary forcing function $\rho(\alpha)$. In this paper we chose to present the result with $\rho(\alpha) = c\alpha^2$ since i) the complexity guarantees do not improve by considering a different forcing function, and because ii) the results and proofs become a bit less transparent. For a general forcing function, the statement would say that if $f(x) - \rho(\alpha) < f(x + \alpha d)$ for all $d \in D$, then

$$\|\nabla f(x)\| \leq \mu^{-1} \left(\frac{L}{2} \alpha d_{max} + \frac{\rho(\alpha)}{\alpha} \frac{1}{d_{min}} \right),$$

where $d_{min} = \min\{\|d\| : d \in D\}$ and $d_{max} = \max\{\|d\| : d \in D\}$. In this general form the lemma is presented, for instance, in [7].

Proof. Since f is L -smooth, (16) implies that for all $x, y \in \mathbb{R}^n$ we have $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$. By assumption, we know that for all $d \in D$, we have $-f(x + \alpha d) < -f(x) + c\alpha^2$. Summing these two inequalities, and setting $y = x + \alpha d$, we obtain

$$0 < \langle \nabla f(x), \alpha d \rangle + c\alpha^2 + \frac{L}{2} \|\alpha d\|^2. \quad (18)$$

Let $d \in D$ be such that $\mu \|\nabla f(x)\| \|d\| = \mu \|\nabla f(x)\| \leq -\langle \nabla f(x), d \rangle$ (see (14)). Inequality (17) follows by multiplying this inequality by α , adding it to (18) and rearranging the result. \square

6.2 Nonconvex case

In this section, we state our most general complexity result – one that does not require any additional assumptions on f , besides smoothness and boundedness. In particular, it applies to non-convex objective functions.

Theorem 11 (Nonconvex case). *Let Assumptions 9 (f is L -smooth) and 8 (D is a positive spanning set) be satisfied. Choose initial iterate $x_0 \in \mathbb{R}^n$ and initial stepsize parameter $\alpha_0 > 0$. Then the iterates $k \geq 1$ of Algorithm 2 satisfy:*

$$\|\nabla f(x_k)\| \leq \frac{(\frac{L}{2} + c) \alpha_0}{\mu 2^k}. \quad (19)$$

Pick any $0 < \epsilon < (\frac{L}{2} + c) \frac{\alpha_0}{\mu}$. If, moreover, Assumption 1 (f is bounded below) is satisfied and we set

$$k \geq k(\epsilon) \stackrel{\text{def}}{=} \left\lceil \log_2 \left(\frac{(\frac{L}{2} + c) \alpha_0}{\mu \epsilon} \right) \right\rceil, \quad (20)$$

then $\|\nabla f(x_{k(\epsilon)})\| \leq \epsilon$, while the method performs in total at most

$$N(k(\epsilon)) \leq 1 + |D| \left(k(\epsilon) + \frac{16(f(x_0) - f^*)(\frac{L}{2} + c)^2}{3c\mu^2\epsilon^2} \right) \quad (21)$$

function evaluations.

Proof. Inequality (19) follows from (17) by construction of x_k (see (4)) and α_k . Since $k(\epsilon) \geq 1$,

$$\|\nabla f(x_{k(\epsilon)})\| \stackrel{(19)}{\leq} \frac{(\frac{L}{2} + c) \alpha_0}{\mu 2^{k(\epsilon)}} \stackrel{(20)}{\leq} \epsilon.$$

By substituting the second estimate in (5) into (6) and using the fact that $\alpha_k = \alpha_0/2^k$, $k \geq 0$, we obtain the bound

$$N(k) \leq 1 + k|D| + \frac{4(4^k - 1)|D|}{3c\alpha_0^2} (f(x_0) - f^*), \quad (22)$$

from which with $k = k(\epsilon)$ we get (21). \square

We shall now briefly comment the above result.

- Notice that we do not enforce Assumption 3 – no initialization is needed if f is not convex. This is because in the nonconvex case the best bound on l_1 is the one we have used in the analysis (given by (5)) – and it is not improved by enforcing Assumption 3. As we shall see, in the convex and strongly convex cases a better bound on l_1 is available if we enforce Assumption 3 (and, as we have seen in Section 4, initialization is cheap), which leads to better complexity.
- In the algorithm we have freedom in choosing c . It is easy to see that the choice $c = \frac{L}{2}$ minimizes the dominant term in the complexity bound (21), in which case the bound takes the form

$$\mathcal{O}\left(\frac{|D|}{\mu^2} \frac{L(f(x_0) - f^*)}{\epsilon^2}\right). \quad (23)$$

Needless to say, in a derivative-free setting the value of L is usually not available and hence usually one cannot choose $c = \frac{L}{2}$. For $c = O(1)$, the complexity depends quadratically on L .

- If D is chosen to be the “maximal positive basis” (see (15)), the bound (23) reduces to

$$\mathcal{O}\left(\frac{n^2 L(f(x_0) - f^*)}{\epsilon^2}\right).$$

This form of the result was used in Table 1.

6.3 Convex case

In this section, we analyze the method under the additional assumption that f is convex. For technical reasons, we also assume that the problem is solvable (i.e., that it has a minimizer x_*) and that, given an initial iterate $x_0 \in \mathbb{R}^n$, the quantity

$$R_0 \stackrel{\text{def}}{=} \sup_{x \in \mathbb{R}^n} \{\|x - x_*\| : f(x) \leq f(x_0)\} \quad (24)$$

is finite. Further, for convenience we define

$$B \stackrel{\text{def}}{=} \frac{R_0(\frac{L}{2} + c)}{\mu}. \quad (25)$$

We are now ready to state the complexity result.

Theorem 12 (Convex case). *Let Assumptions 9 (f is L -smooth) and 8 (D is a positive spanning set) be satisfied. Further assume that f is convex, has a minimizer x_* and $R_0 < \infty$ for some initial iterate $x_0 \in \mathbb{R}^n$. Finally, let Assumption 3 (initialization) be satisfied. Then:*

(i) *The iterates $\{x_k\}_{k \geq 0}$ of Algorithm 2 satisfy*

$$\|\nabla f(x_k)\| \leq \frac{(\frac{L}{2} + c)\alpha_0}{\mu 2^k}, \quad f(x_k) - f(x_*) \leq \frac{B\alpha_0}{2^k}, \quad k \geq 0, \quad (26)$$

where at iteration k the method needs to perform at most $|D| \left(1 + \frac{2^{k+1}B}{c\alpha_0}\right)$ function evaluations.

(ii) In particular, if we set $k = k(\epsilon) \stackrel{\text{def}}{=} \lceil \log_2 \left(\frac{B\alpha_0}{\epsilon} \right) \rceil$, where $0 < \epsilon \leq B\alpha_0$, then $f(x_k) - f(x_*) \leq \epsilon$, while Algorithm 2 performs in total at most

$$N(k(\epsilon)) \leq 1 + |D| \left(k(\epsilon) + \frac{8B^2}{c\epsilon} \right) \quad (27)$$

function evaluations.

Proof. The first part of (26), for $k \geq 1$, follows from Theorem 11. For $k = 0$ it follows by combining Assumption 3 and Lemma 10. In order to establish the second part of (26), it suffices to note that since $f(x_k) \leq f(x_0)$ for all k (see Lemma 2(i)), we have for all $k \geq 0$:

$$f(x_k) - f(x_*) \leq \langle \nabla f(x_k), x_k - x_* \rangle \leq \|\nabla f(x_k)\| \|x_k - x_*\| \stackrel{(24)}{\leq} \|\nabla f(x_k)\| R_0 \stackrel{(26)+(25)}{\leq} B\alpha_k. \quad (28)$$

It only remains to establish (27). Letting $r_k = f(x_k) - f^*$, and using (3) and (28), we have $0 \leq r_k \leq r_{k-1} - l_k c \alpha_k^2 \leq B\alpha_{k-1} - l_k c \alpha_k^2 = 2B\alpha_k - l_k c \alpha_k^2$, whence

$$l_k \leq \frac{2B}{c\alpha_k} = \frac{2^{k+1}B}{c\alpha_0}. \quad (29)$$

We can now estimate the total number of function evaluations by plugging (37) into (6):

$$\begin{aligned} N(k(\epsilon)) &\stackrel{(6)}{=} 1 + \sum_{k=1}^{k(\epsilon)} |D|(l_k + 1) \stackrel{(37)}{\leq} 1 + |D| \sum_{k=1}^{k(\epsilon)} \left(\frac{2B}{c\alpha_k} + 1 \right) = 1 + |D|k(\epsilon) + \frac{2B|D|}{c\alpha_0} \sum_{k=1}^{k(\epsilon)} 2^k \\ &\leq 1 + |D|k(\epsilon) + \frac{2B|D|}{c\alpha_0} 2^{k(\epsilon)+1} \leq 1 + |D| \left(k(\epsilon) + \frac{8B^2}{c\epsilon} \right). \end{aligned}$$

□

We shall now comment the result.

- Note that Assumption 3 was used to argue that $r_0 \leq B\alpha_0$, which was in turn used to bound l_1 . The bounds on l_k for $k > 1$ hold even without this assumption. Alternatively, we could have skipped Assumption 3 and bounded l_1 as in Theorem 11. Relations (26) would hold for $k \geq 1$.
- Again, we have freedom in choosing c (and note that c appears also in the definition of B). It is easy to see that the choice $c = \frac{L}{2}$ minimizes the dominating term $\frac{B^2}{c}$ in the complexity bound (27), in which case $B = \frac{LR_0}{\mu}$, $\frac{B^2}{c} = \frac{2LR_0^2}{\mu^2}$ and the bound (27) takes the form

$$1 + |D| \left[\left\lceil \log_2 \left(\frac{LR_0\alpha_0}{\mu\epsilon} \right) \right\rceil + \frac{16LR_0^2}{\mu^2\epsilon} \right] = \mathcal{O} \left(\frac{|D|}{\mu^2} \frac{LR_0^2}{\epsilon} \right). \quad (30)$$

- If D is chosen to be the “maximal positive basis” (see (15)), the bound (30) reduces to

$$\mathcal{O} \left(\frac{n^2 LR_0^2}{\epsilon} \right).$$

The result is listed in this form in Table 1.

- It is possible to improve the algorithm by introducing an additional stopping criterion: $l_k \geq \frac{B}{c\alpha_k}$. The analysis is almost the same, and the resulting number of function evaluations is halved in this case. However, this improvement is rather theoretical, since we typically do not know the value of B .

6.4 Strongly convex case

In this section we introduce an additional assumption: f is λ -strongly convex for some (strong convexity) constant $\lambda > 0$. That is, we require that $\forall x, y \in \mathbb{R}^n$, we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\lambda}{2} \|y - x\|^2. \quad (31)$$

In particular, by minimizing both sides of the above inequality in y , we obtain the standard inequality

$$f(x) - f(x_*) \leq \frac{1}{2\lambda} \|\nabla f(x)\|^2. \quad (32)$$

Moreover, by substituting $y \leftarrow x$ and $x \leftarrow x_*$ into (31), and using the fact that $\nabla f(x_*) = 0$, we obtain another well known inequality:

$$\frac{\lambda}{2} \|x - x_*\|^2 \leq f(x) - f(x_*). \quad (33)$$

To simplify notation, in what follows we will make use of the following quantity:

$$S \stackrel{\text{def}}{=} \frac{(\frac{L}{2} + c)^2}{2\lambda\mu^2}. \quad (34)$$

Theorem 13 (Strongly convex case). *Let Assumptions 9 (f is L -smooth), 8 (D is a positive spanning set) and 3 (initialization) be satisfied. Further, assume that f is λ -strongly convex. Then:*

(i) *The iterates $\{x_k\}_{k \geq 0}$ of Algorithm 2 satisfy*

$$\|\nabla f(x_k)\| \leq \frac{(\frac{L}{2} + c)\alpha_0}{2^k \mu}, \quad f(x_k) - f(x_*) \leq S \left(\frac{\alpha_0}{2^k} \right)^2, \quad \|x_k - x_*\| \leq \frac{(\frac{L}{2} + c)\alpha_0}{2^k \mu \lambda}, \quad (35)$$

where at each iteration the method needs to perform at most $|D| \left(\frac{4S}{c} + 1 \right)$ function evaluations.

(ii) *In particular, if we set $k = k(\epsilon) \stackrel{\text{def}}{=} \left\lceil \log_2 \left(\alpha_0 \sqrt{\frac{S}{\epsilon}} \right) \right\rceil$, where $0 < \epsilon < S\alpha_0^2$, then $f(x_k) - f(x_*) \leq \epsilon$, while the method performs in total at most*

$$N(k(\epsilon)) \leq 1 + |D| \left(\frac{4S}{c} + 1 \right) \left(1 + \log_2 \left(\alpha_0 \sqrt{\frac{S}{\epsilon}} \right) \right) \quad (36)$$

function evaluations.

Proof. The first part of (35) was already proved in the convex case (26); the rest follows from

$$f(x_k) - f(x_*) \stackrel{(32)}{\leq} \frac{1}{2\lambda} \|\nabla f(x_k)\|^2 \stackrel{(35)+(34)}{\leq} S\alpha_k^2; \quad \|x_k - x_*\| \stackrel{(33)+(32)}{\leq} \frac{\|\nabla f(x_k)\|}{\lambda} \stackrel{(35)}{\leq} \frac{(\frac{L}{2} + c)\alpha_0}{2^k \mu \lambda}.$$

It only remains to establish the bound (36). Letting $r_k = f(x_k) - f^*$, and using (3) and the second inequality in (35), we have $0 \leq r_k \leq r_{k-1} - l_k c \alpha_k^2 \leq S(2\alpha_k)^2 - l_k c \alpha_k^2$, whence

$$l_k \leq \frac{4S\alpha_k^2}{c\alpha_k^2} = \frac{4S}{c}. \quad (37)$$

We can now estimate the total number of function evaluations by plugging (37) into (6):

$$N(k(\epsilon)) \stackrel{(6)}{=} 1 + \sum_{k=1}^{k(\epsilon)} |D|(l_k + 1) \stackrel{(37)}{\leq} 1 + \sum_{k=1}^{k(\epsilon)} |D| \left(\frac{4S}{c} + 1 \right) = 1 + |D| \left(\frac{4S}{c} + 1 \right) k(\epsilon).$$

□

Let us now comment on the result.

- As before, in the algorithm we have freedom in choosing c . Choosing $c = \frac{L}{2}$ minimizes the dominating term $\frac{S}{c}$ in the complexity bound (36), in which case $S = \frac{L^2}{2\lambda\mu^2}$, $\frac{S}{c} = \frac{L}{\lambda\mu^2}$ and the bound (36) takes the form

$$1 + |D| \left(1 + \frac{4L}{\lambda\mu^2} \right) \left(1 + \log_2 \left(\frac{\alpha_0 L}{\mu} \sqrt{\frac{1}{2\lambda\epsilon}} \right) \right). \quad (38)$$

- If D is chosen to be the “maximal positive basis” (see (15)), the bound (38) reduces to

$$\mathcal{O} \left(\frac{n^2 L}{\lambda} \log_2 \left(\frac{n L^2 \alpha_0^2}{\lambda \epsilon} \right) \right) = \tilde{\mathcal{O}} \left(n^2 \frac{L}{\lambda} \right),$$

where the $\tilde{\mathcal{O}}$ notation suppresses the logarithmic term. The complexity is proportional to the condition number L/μ .

- As in the convex case, we can introduce the additional stopping criterion $l_k \leq \frac{3S}{c}$. The analysis is similar and the bound on function evaluation can be reduced by the factor of 4/3. However, in practice we often do not know S .

7 Conclusion

We proposed and analyzed the complexity of SDS – a simplified variant of the direct search method. Thanks to the simplified design of our method, and two novel and efficient initialization strategies, our method depends on a single parameter only. This contrasts with standard direct search which depends on a large number of parameters. We gave the *first unified analysis*, covering three classes of unconstrained smooth minimization problems: non-convex, convex and strongly convex. The analysis follows the same pattern in all three cases. Finally, our complexity bounds have a simple form, are easy to interpret, and are better than existing bounds in the convex case.

References

- [1] M A Abramson and C Audet. Convergence of mesh adaptive direct search to second-order stationary points. *SIAM Journal on Optimization*, 17:606–619, 2006.
- [2] C Audet and J E Dennis Jr. Analysis of generalized pattern searches. *SIAM Journal on Optimization*, 13(3):889–903, 2002.

- [3] C Audet and J E Dennis Jr. Mesh adaptive direct search algorithms for constrained optimization. *SIAM Journal on Optimization*, 17(1):188–217, 2006.
- [4] C Audet and D Orban. Finding optimal algorithmic parameters using derivative-free optimization. *SIAM Journal on Optimization*, 17:642–664, 2006.
- [5] C Cartis, N I M Gould, and P L Toint. On the complexity of steepest descent, Newton’s and regularized Newton’s methods for nonconvex unconstrained optimization problems. *SIAM Journal on Optimization*, 20(6):2833–2852, 2010.
- [6] C Cartis, N I M Gould, and P L Toint. On the oracle complexity of first-order and derivative-free algorithms for smooth nonconvex minimization. *SIAM Journal on Optimization*, 22(1):66–86, 2012.
- [7] A R Conn, K Scheinberg, and L N Vicente. *Introduction to Derivative-Free Optimization*. MPS-SIAM Series on Optimization. SIAM, 2009.
- [8] M Dodangeh and L N Vicente. Worst case complexity of direct search under convexity. Technical report, Technical Report 13-10, Dept. Mathematics, Univ. Coimbra, 2013.
- [9] E D Dolan, R M Lewis, and V Torczon. On the local convergence of pattern search. *SIAM Journal on Optimization*, 14(2):567–583, 2003.
- [10] J C Duchi, M I Jordan, M J Wainwright, and A Wibisono. Optimal rates for zero-order optimization: the power of two function evaluations. *arXiv:1312.2139*, 2013.
- [11] R Garmanjani and L N Vicente. Smoothing and worst case complexity for direct-search methods in non-smooth optimization. *IMA Journal of Numerical Analysis*, 2012.
- [12] S Ghadimi and G Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
- [13] S Gratton, C W Royer, L N Vicente, and Z Zhang. Direct search based on probabilistic descent. Technical report.
- [14] R Hooke and T A Jeeves. Direct search solution of numerical and statistical problems. *Journal of the ACM*, 8(2):212–229, 1961.
- [15] T G Kolda, R M Lewis, and V Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review*, 45:385–482, 2003.
- [16] Yu Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publishers, 2004.
- [17] Yu Nesterov. Random gradient-free minimization of convex functions. Technical report, Université catholique de Louvain, CORE Discussion Paper 2011/16, 2011.
- [18] B Recht, K G Jamieson, and R Nowak. Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2012.
- [19] P Richtárik and M Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

- [20] V Torczon. On the convergence of pattern search algorithms. *SIAM Journal on Optimization*, 7:1–25, 1997.
- [21] L N Vicente. Worst case complexity of direct search. *EURO Journal on Computational Optimization*, 1(1-2):143–153, 2013.
- [22] L N Vicente and A L Custódio. Analysis of direct searches for discontinuous functions. Technical report, 2010.
- [23] Yu Wen-ci. Positive basis and a class of direct search techniques. *Scientia Sinica, Special Issue of Mathematics*, 1:53–67, 1979.

Appendix: Proof Sketch of Proposition 7

Proposition 7. *Let D be a finite set of nonzero vectors. Then the following statements are equivalent:*

- (i) D is a positive spanning set.
- (ii) The cosine measure of D is positive:

$$\mu \stackrel{\text{def}}{=} \min_{0 \neq v \in \mathbb{R}^n} \max_{d \in D} \frac{\langle v, d \rangle}{\|v\| \|d\|} > 0. \quad (39)$$

Above, $\langle \cdot, \cdot \rangle$ is the standard Euclidean inner product and $\|\cdot\|$ is the standard Euclidean norm.

- (iii) The convex hull of D contains 0 in its interior.

Proof. (i) \Leftrightarrow (ii). Equivalence of (i) and (ii) can be shown by a separation argument.

(ii) \Rightarrow (iii). Let \mathcal{B} be the unit Euclidean ball in \mathbb{R}^n centered at the origin. Note that 0 is in the interior of $\text{Conv}(D)$ if and only if it is in the interior of $\mathcal{D} \stackrel{\text{def}}{=} \text{Conv}(D')$, where $D' = \{d/\|d\| : d \in D\}$. Further, note that in view of (39), for every $0 \neq v$ we have

$$\sigma_{\mathcal{D}}(v) \stackrel{\text{def}}{=} \max_{d' \in \mathcal{D}} \langle v, d' \rangle = \max_{d' \in D'} \langle v, d' \rangle = \max_{d \in D} \frac{\langle v, d \rangle}{\|d\|} \stackrel{(39)}{\geq} \mu \|v\|.$$

On the other hand, for any $\nu > 0$,

$$\sigma_{\nu\mathcal{B}}(v) \stackrel{\text{def}}{=} \max_{d' \in \nu\mathcal{B}} \langle v, d' \rangle = \nu \|v\|.$$

If (ii) holds, then $\mu > 0$, and hence $\sigma_{\mathcal{D}} \geq \sigma_{\mu\mathcal{B}}$, implying that $\mathcal{D} \supset \mu\mathcal{B}$, establishing (iii).

(iii) \Leftrightarrow (i). If (iii) holds, then $\text{Conv}(D) \supset \nu\mathcal{B}$ for some $\nu > 0$. Let $\text{Cone}(D)$ be the conic hull of D (the smallest convex cone containing D). Since $\text{Cone}(D) \supset \text{Conv}(D) \supset \nu\mathcal{B}$, we must have $tb \in \text{Cone}(D)$ for all $b \in \mathcal{B}$ and $t \geq 0$. That is, $\text{Cone}(D) = \mathbb{R}^n$, implying (i). \square